View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

PERGAMON

Vision Research 43 (2003) 393–404

Vision Research

www.elsevier.com/locate/visres

Representation of statistical properties

Sang Chul Chong *, Anne Treisman

Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544, USA

Received 13 May 2002; received in revised form 13 November 2002

Abstract

Everyday scenes often contain sets of similar objects. Perceptual representations may summarize these with statistical descriptors. After determining the psychological mean of two sizes, we measured thresholds for judging the mean with arrays of 12 circles of heterogeneous sizes. They were close to those for the size of elements in homogeneous arrays and single elements, and were little affected by either exposure duration (50–1000 ms) or memory delays (up to 2s). They were only slightly more accurate within the same distribution than across different distributions (normal, uniform, two-peaks, and homogeneous), confirming that subjects were indeed averaging sizes.

© 2003 Elsevier Science Ltd. All rights reserved.

Keywords: Statistical properties; Size; Perception; Mean

As we move around the environment, we feel that we are seeing a complete and veridical perceptual representation of the surrounding scene, akin to a high resolution, full-color photograph. How can we achieve this impression, when acuity and color sensitivity rapidly drop off with distance from the fixation point? Historically, the answer has been the composite image hypothesis (Davidson, Fox, & Dick, 1973). According to this hypothesis, the visual system builds up a composite perceptual image over consecutive fixations by overlapping successive perceptual images in a system that maps a retinal reference frame onto a spatiotopic reference frame. However, psychophysical and behavioral data have almost uniformly provided evidence against this hypothesis. Irwin (1991) showed that when two dot patterns forming a matrix of dots are presented in rapid succession at the same spatial position within a single fixation, a fused pattern is perceived. However, if a saccade is made between the first and second patterns, no perceptual fusion occurs. It seems unlikely, then, that we build up a composite perceptual image across saccades by spatially aligning information from each fixation. It seems more likely that participants abstract a schematic representation of a scene from several successive fixations (Hochberg, 1978; Hock & Schmelzkopf, 1980). However, the nature of the schematic representation is still unclear.

Change detection experiments also cast doubt on the introspective impression of a rich and detailed representation. In these experiments, an original and a modified image are presented in rapid alternation with a blank screen between them. Observers have considerable difficulty in detecting even major changes in alternating scenes unless they are directly attending to the changing object (Rensink, O'Regan, & Clark, 1997).

The visual world is highly redundant. Most surfaces have fairly uniform properties with only occasional discontinuities. Many elements and objects are replicated within neighboring areas, for example the leaves on a tree, the cars in a car park, a flock of flying birds. Statistical properties, such as the mean, range and variance of the size, color, orientation, or speed and direction of motion of elements in the display may play a part in forming schematic perceptual representations. We can discriminate subtle color differences between individual leaves if we attend to them, but otherwise we register and retain just the global impression of variegated greens on the tree as a whole. Ariely (2001) and Ariely and Burbeck (1995) proposed that the visual system represents overall statistical properties when sets of similar objects are present. The apparently complete and veridical perceptual representation of the surrounding scene that we experience may be an illusion generated from occasional detailed samples together

* Corresponding author.
E-mail address: scchong@princeton.edu (S.C. Chong).

0042-6989/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
PII: S0042-6989(02)00596-5

with statistical summaries of remaining areas and an overall interpretation of the meaning or gist. If this is the case, it should be important to study how the statistical properties are encoded and represented.

In motion perception, our ability to use statistical properties is impressive. Given a stimulus containing many different local motion directions, we form a unified global percept of motion in the direction of the mean (Williams & Sekuler, 1984). We can discriminate between such global percepts when they differ by as little as 1°–2° for distributions containing up to about 45 different directions (Watamaniuk, Sekuler, & Williams, 1989). The visual system can also average speed information. Watamaniuk and Duchon (1992) found that participants based their discrimination of speed on the mean speed of the stimulus, with average speed-discrimination thresholds ranging from 5–10%, comparable to those obtained with stimuli in which all dots move at the same speed (De Bruyn & Orban, 1988; Snowden & Braddick, 1991).

Statistics are also perceptually available in the domain of orientation. Participants are highly accurate at performing mean orientation judgments. Thresholds are as low as 1.5° for line textures, 2.5° for Glass patterns (Dakin, 1997) and 1.2°–2.5° for Gaussian distributed orientations (Dakin & Watt, 1997), comparable to orientation thresholds reported for single line and grating stimuli (Heeley & Buchanan-Smith, 1990). Observers can even reliably estimate the average orientation of crowded Gabor patches when these are presented peripherally and too crowded to allow the discrimination of individual orientations (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001).

Sensory neurons appear to have adapted, through both evolutionary and developmental processes, to match the statistical properties of the signals to which they are exposed (Simoncelli & Olshausen, 2001). Barlow (1961) proposed that information theory could provide a link between environmental statistics and neural responses, suggesting that the role of early sensory neurons is to remove statistical redundancy in the sensory input. Consistent with this suggestion, individual neurons rapidly adapt to changes in contrast and spatial scale (Smirnakis, Berry, Warland, Bialek, & Meister, 1997), orientation (Müller, Metha, Krauskopf, & Lennie, 1999), and variance of velocity (Brenner, Bialek, & de Ruyter van Steveninck, 2000).

In the present paper, we explore the evidence for statistical processing in the domain of size, and attempt to measure it directly. The starting point was a finding by Ariely (2001) and Ariely and Burbeck (1995), who showed that participants are considerably better at judging the mean size of a set of circles than at judging the size of any randomly selected member of the set. Ariely presented displays of circles of various sizes. In the mean judgment task, these were followed by a single probe circle to be judged as larger or smaller than the mean. In the member identification task, the display was followed in one experiment by a single probe circle to be judged as having been present or absent in the preceding display, and in another experiment by a pair of circles for a forced choice judgment of which had been present in the preceding display. Note that these tests depended on immediate memory for the display.

By asking which of two displays had the larger mean, our experiments compared discrimination when both displays were present together to performance with successive presentation at ISIs of either 100 ms or 2 s. Thus we could compare immediate perception with memory and memory decay, if any. We also compared perception of the mean with perception of individual sizes, using three kinds of size judgments: judgments of the mean size in heterogeneous displays, judgments of the same-sized items in homogeneous displays, and judgments of the size of single items presented alone. In subsequent experiments we explored the effects on mean size judgments of varying the exposure duration, and the efficiency of statistical judgments of the mean size within sets drawn from the same distribution or across sets drawn from different distributions.

Before testing perception or memory for the mean size of sets of circles, it seemed important to determine what is in fact perceived as the mean size, using just two items. Is it the arithmetic mean of the diameters, or of the areas, or should we use a logarithmic scale, as Weber's law might suggest, or a power function, which, according to Teghtsoonian (1965), gives the best estimates of size perception using a magnitude estimation procedure (Stevens, 1957). We also investigated whether estimates of the mean size differ for one and for two-dimensional stimuli, comparing lines and circles. The details of this experiment are given in Appendix A.

The method and results can be summarized as follows. Participants saw two circles (or two lines) in the upper half of the display and were asked to adjust the size of a third circle (or line) in the lower half of the display to match the mean size of the two presented stimuli. The initial size of the adjustable stimulus was either small (3.60°–5.01°) or large (15.89°–14.48°). The participants served in one block testing perception, in which the two fixed stimuli remained present while the adjustment was made, and one testing memory, in which the two fixed stimuli were presented for 1 s only. Each block was preceded by two practice trials.

We report only the results for the circles here. The mean size estimates were the same for the perception and the memory blocks although the variance was larger for memory. Participants' estimates differed significantly from the geometric mean ($t_{(25)} = 16.315$, $p < 0.01$), the arithmetic mean of the diameters ($t_{(25)} = 4.762$, $p < 0.01$), and the arithmetic mean of the areas ($t_{(25)} = -5.514$, $p < 0.01$). The results approximated the power

function with an exponent of 0.76 previously reported by Teghtsoonian (1965) using the method of magnitude-estimation for judgments of the size of a single stimulus rather than the mean of two. Note that the power function with the exponent of 0.76 predicts a mean that lies between the mean of the areas and the mean of the diameters. One possible explanation is that participants divided their estimates between matching the mean area and matching the mean diameter length. The values are too close for our data to distinguish whether the participants could be divided into two groups, one matching each of those criteria. The results give us the information we need to interpret the results of subsequent experiments and to assess participants' ability to extract the mean of displays containing more than two circles.

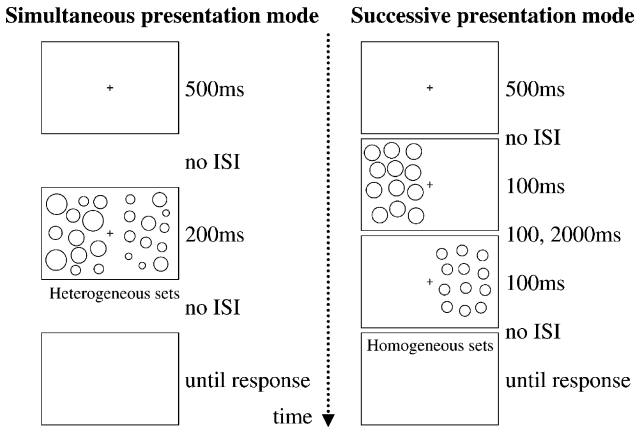

Fig. 1. The timelines for Experiment 1. (a) Examples of the timeline of the simultaneous presentation mode and of heterogeneous stimulus sets. (b) Examples of the timeline of the successive presentation mode and of homogeneous stimulus sets.

1. Experiment 1

In Experiment 1, we measured thresholds for judgments of the mean size of 12 circles of varied sizes, using the method of constant stimuli. We compared these to thresholds for judgments of the sizes of a set of 12 identical circles in a display, and for judgments of the size of a single circle presented alone.

Ariely (2001) and Ariely and Burbeck (1995) found that judgments of the mean size in heterogeneous arrays were more accurate than judgments of individual member sizes in the same arrays. In fact his participants proved quite unable to discriminate between specific items randomly selected from the display and new items that were within the same range of sizes but that had not been presented. His goal was to see whether the ability to identify the mean size of a set depends on the ability to identify the individual elements of the set. His surprising conclusion was that the mean judgment was a separate and much more efficient process. The purpose of our Experiment 1 was to test just how accurately we could judge the mean size of a set, and to compare these judgments with the accuracy of judging the size of single items presented alone and judgments of homogeneous sets of items. We also tested how these abilities were affected by different time delays.

1.1. Method

1.1.1. Participants

Five participants including the first author participated in the experiment. All were members of Princeton University. All had normal or corrected-to-normal vision.

1.1.2. Apparatus and stimuli

The stimuli were presented on the screen of a Samsung SyncMaster 955DF 19 in. Monitor. The monitor was driven by a Macintosh G4, which also performed all timing functions and controlled the course of the experiment. Participants viewed the screen with both eyes and were seated approximately 66 cm from the screen. The stimuli are shown in Fig. 1. Each display was divided into two halves vertically, each containing either 1 or 12 circles in either one or a mixture of four sizes. The sizes were equally spaced on a log scale separated by a factor of 1.25. [1] The mean circle diameter was 2.63° and the diameters ranged from 1.82° to 3.56°. The left and right displays were separated by 6.32° in their near edges. Each visual field had an imaginary 4×4 matrix where each cell measured $6.32° \times 6.32°$. The locations of the circles within the displays were randomly selected in the matrix and they were randomly jittered within the range of 0.49° in each cell of the matrix. When only one circle was presented in each visual field, it was always presented in the center of the matrix. In each trial all of the circles shown were randomly scaled by a small multiplicative factor to discourage the participants from basing their judgments on previously seen stimuli. Four multiplicative factors (0.7, 0.8, 0.9, 1) were used and the same factor scaled all circles in any one trial. The luminance of the stimuli was 49.93 cd/m^2 and the luminance of the black background was 0.006 cd/m^2.

1.1.3. Design

The task was to say which side of the display had the larger size or the larger mean size. There were two independent variables in the experiment, which were both

[1] This experiment was actually run before the pilot study described above, or we would have used the power function rather than a log scale. However, there were only slight differences between the arithmetic mean of the diameters and the mean of the power function values. These differences disappeared in the actual stimuli because all the differences were less than one pixel.

varied within participants. The first variable was the type of size comparison to be made between the left and the right array—either the mean sizes of the heterogeneous arrays, or the sizes of the circles in the two homogeneous arrays, or the sizes of two single circles presented alone. The second variable was the presentation mode—either simultaneous, or successive. With successive presentations, 2 ISIs were tested, 100 ms and 2 s.

Each participant served in at least four sessions containing six blocks each (3 types of size discrimination × 2 presentation modes) as well as six practice blocks. The discrimination type (heterogeneous, homogeneous, and single) and presentation mode (simultaneous or successive) were blocked and the order of blocks was counterbalanced within and across participants. The two ISIs in the successive presentation condition were randomly mixed within the successive presentation blocks. There were 21 trials in the practice blocks, 96 trials (6 comparison stimuli × 16 repetitions) in the simultaneous presentation condition, and 192 trials (2 ISIs × 6 comparison stimuli × 16 repetitions) in the experimental blocks of the successive presentation condition. The order of trials within each block was randomly selected under the constraint that each condition (comparison stimuli or ISI) was presented once before any condition was repeated.

Thresholds were measured using the method of constant stimuli in which participants decided on each trial which visual field had the larger size or the larger mean size. The circles on each side differed by a constant difference in diameter within any given display. There were six constant differences between the two displays, 2%, 4%, 6%, 8%, 10% and 12% diameter difference on the power function scale. An equal number of trials with each constant difference were randomly mixed in the experiment. Probit analysis (Finney, 1971) was used to determine the thresholds. This procedure plots the proportion of correct judgments against each difference between the two displays. The threshold was defined as the percent diameter difference between the two displays that gave 75% accuracy in this graph. When we could not decide the threshold due to low accuracy, we reran that block with a wider range of stepwise differences. Only one participant needed an extra step of 14% diameter difference for the successive presentation mode with both 100 ms and 2 s delay.

1.1.4. Procedure

A timeline of the procedure is shown in Fig. 1. Each trial started with a fixation cross for 500 ms. In the simultaneous presentation condition, 12 circles of 4 different sizes, 12 circles of the same size, or an individual circle were presented at the same time for 200 ms in each visual field. In the successive presentation condition, the circles in the left visual field were presented first for 100 ms and the circles in the right visual field were presented for 100 ms either 100 ms or 2 s later. Participants' task was to decide either which visual field had the larger mean size or which visual field had the larger size. When they thought that the left visual field had either the larger mean size or the larger size, they pressed '1'. When they thought that the right visual field had either the larger mean size or the larger size, they pressed '2'. When their decision was incorrect, they heard a short high-pitched tone.

1.2. Results and discussion

The results of Experiment 1 are shown in Fig. 2. The thresholds were low for all three types of size judgment. A diameter difference of only 6–8% was required for 75% accuracy in mean judgments when the stimuli were presented simultaneously. Delays of up to 2 s had little effect on the thresholds for the homogeneous arrays of circles. However, the thresholds for the heterogeneous arrays and the single circles did increase with delay.

An ANOVA indicated significant effects of discrimination type ($F_{(2,32)} = 8.591$, $p < 0.01$) and of presentation delay ($F_{(2,32)} = 13.284$, $p < 0.01$). According to Bonferroni post hoc analysis, these differences were due to significantly higher thresholds with heterogeneous than with homogeneous displays, and significantly higher thresholds at 2 s than at 100 ms or 0 ms delays. The interaction between the type of size judgment and the presentation delay was not significant ($F_{(2,32)} = 1.53$, $p = 0.22$). However, separate analyses of the effect of size judgment type for each presentation mode revealed that the homogeneous condition was different from the mean and single item conditions at 2 s delay ($F_{(2,8)} = 11.238$, $p < 0.01$), but there were no significant effects of

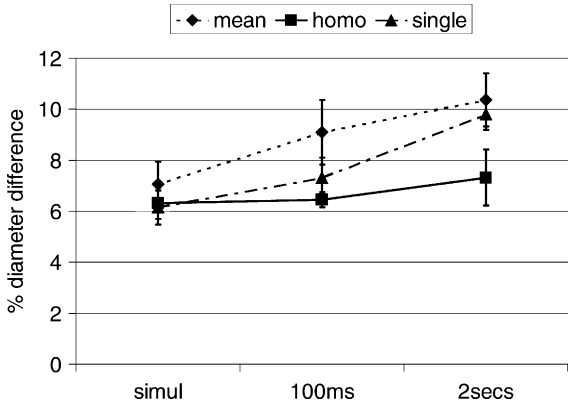

Fig. 2. The results of Experiment 1. The *Y*-axis indicates the thresholds defined as the percent diameter difference between the two displays on any given trial. The *X*-axis indicates the delays between the two displays and SIMUL stands for simultaneous presentation. The error bars indicate the standard errors.

size judgment type at 100 ms delay ($F_{(2,8)} = 3.624$, $p = 0.08$) or in the simultaneous condition ($F_{(2,8)} = 1.148$, $p = 0.36$).

The thresholds for mean size in our experiment were similar to those found by Ariely (2001) and Ariely and Burbeck (1995). The size differences in our set were between those in the two sets used by Ariely. Ours were separated by a factor of 1.25, giving a threshold of 8–10% in the delay conditions, whereas his scaling factors were 1.05 in his similar set, giving a threshold of 4–6%, and 1.4 in his dissimilar set, giving a threshold of 6–12% with successive presentation. Performance on the single items was much better in our experiment than in Ariely's. This is not surprising since in our experiments, comparison of two single items were made on single item displays, so that attention could be focused on the two relevant items. In Ariely's experiment, the single item was sampled after the presentation from a multi-item display.

Our finding that the comparisons of mean size were as accurate as comparisons of two single items is quite surprising. With an exposure duration of 200 ms, it is unlikely that participants had time to calculate the mean size by adding each size and then dividing the sum by the total number of circles. This suggests that the process of extracting the mean size might be a parallel preattentive process. Its limits are tested in the next experiment where we vary the exposure duration.

2. Experiment 2

In Experiment 2 we investigated how the exposure duration affected judgments of the mean size of heterogeneous, and homogeneous arrays and of a single pair of circles.

2.1. Methods

2.1.1. Participants

The same five participants as in Experiment 1 were tested in this experiment.

2.1.2. Apparatus and stimuli

The stimuli and the luminance were the same as in Experiment 1 except that a different monitor and a different computer were used. The stimuli were presented on the screen of an Apple 17 in. Monitor, which was driven by a Macintosh G3. Participants were seated approximately 66 cm from the screen. The sizes in this experiment were slightly smaller than those in Experiment 1 because of the smaller monitor. The mean circle diameter was 2.35° and the diameters ranged from 1.63° to 3.18°.

2.1.3. Design

There were two independent variables in the experiment, which were both varied within participants. The first variable was the type of size comparison to be made between the left and the right array—either the mean sizes of the heterogeneous arrays, or the sizes of the circles in the two homogeneous arrays, or the sizes of two single circles presented alone. The second variable was the exposure duration of the stimuli—either 50 ms, 100 ms, or 1 s.

Each participant served in two sessions consisting of three blocks each (three types of size discrimination) as well as three practice blocks. The three stimulus durations were intermixed in each block. There were 21 trials in the practice blocks, 336 trials in the experimental blocks (7 comparison stimuli × 3 exposure durations × 16 repetitions). The order of blocks was counterbalanced within and across participants. The order of trials within each block was randomly selected under the constraint that each condition was presented once before any condition was repeated.

Thresholds were estimated using the same method as in Experiment 1 except that seven comparison stimuli were used with an additional step of 14% diameter difference.

2.1.4. Procedure

The timeline of this experiment's procedure and the task were the same as for the simultaneous presentation condition in Experiment 1 except that the presentation time varied within each block.

2.2. Results and discussion

The results of Experiment 2 are shown in Fig. 3. Overall thresholds differed significantly across the size judgment conditions ($F_{(2,32)} = 7.485$, $p < 0.01$). A Bonferroni post hoc analysis indicated that the threshold in the homogeneous condition was significantly lower than

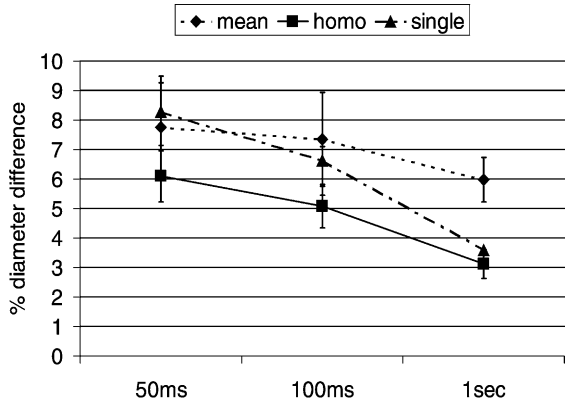

Fig. 3. The results of Experiment 2. The X-axis indicates the duration of the stimuli.

the threshold in the heterogeneous mean condition. The thresholds decreased as the duration was increased ($F_{(2,32)} = 14.889$, $p < 0.01$). A Bonferroni post hoc analysis indicated that the thresholds at 50 and 100 ms durations were significantly higher than those at 1 s duration. The interaction between the type of size judgment and the presentation duration was not significant ($F_{(2,32)} = 1.042$, $p = 0.40$). However, when we looked separately at the effect of size judgment at each presentation duration, the threshold for the mean size was higher than the threshold for the homogeneous and single circle conditions at 1 s duration ($F_{(2,8)} = 9.362$, $p < 0.01$), but there were no significant differences at 100 ms duration ($F_{(2,8)} = 2.076$, $p = 0.19$) or 50 ms duration ($F_{(2,8)} = 4.028$, $p = 0.06$).

It is striking that there was so little deterioration in mean size judgments as the exposure duration was reduced to only 50 ms. It seems that participants are capable of extracting the mean size of two displays of 12 circles each quite accurately in as little as 50 ms. The single item appeared to benefit a little more from the longer exposure duration of 1 s although the interaction did not reach significance. There may be a floor effect on the mean judgments, limiting the improvement that is possible. Internal noise in the averaging process could prevent the increased accuracy that is possible with increased exposure to a single item.

3. Experiment 3

In the final experiment, we tested comparisons of mean size across different distributions of sizes, to see how thresholds for the mean size would be affected. The experiments so far have used a uniform distribution in generating the heterogeneous displays (equal numbers from each of four sizes). If the participants randomly selected one size in a visual field and compared it to a closest match in the opposite visual field, or if they simply compared the largest size across the two displays, they could successfully perform a mean discrimination without averaging any size. To rule out this strategy, we used different distributions in some conditions of Experiment 3, ruling out the option of comparing individual circle sizes. We compared participants' performance in judging mean sizes across different distributions and within the same distribution.

3.1. Method

3.1.1. Participants

The same five participants as in Experiment 1 and an additional two naïve participants were tested in the experiment.

3.1.2. Apparatus and stimuli

The apparatus, the stimuli, and the luminance were the same as in Experiment 2 except that four different distributions were used in Experiment 3. The four different distributions are shown in Fig. 4. The uniform distribution had equal numbers of each of four different sizes (three circles for each of four different sizes). A two-peaks distribution had equal numbers of two different sizes (six instances each of the smallest and the largest circle from the uniform distribution). The normal distribution had unequal numbers of four different sizes (two instances each of the smallest and the largest size and four instances of the two intermediate-sized circles). The homogeneous distribution had only one size (twelve

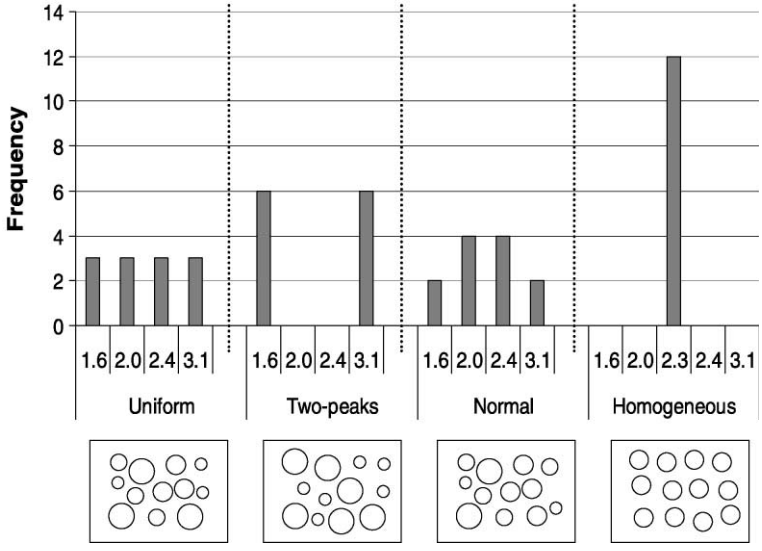

Fig. 4. The four different distributions. The frequency of each size in each type of display, as well as one example of each distribution is shown. The numbers on the X-axis indicate the size of each circle in visual angle.

circles of the mean size of the other distributions). The mean size was the same for all four distributions.

3.1.3. Design

All possible pairs of the four different distributions (10 altogether) were tested with the five experienced participants. The two new participants were tested on the six possible pairs among three distributions (uniform, two-peaks, and homogeneous distributions). All factors were varied within participants.

The five experienced participants served in two sessions of ten blocks each (10 pair-wise comparisons) as well as one practice block. The two new participants served in two sessions of six blocks (6 pair-wise comparisons) as well as one practice block. There were 30 trials in the practice blocks, 112 trials (7 comparison stimuli $\times$ 16 repetitions) in the experimental blocks. The order of blocks was counterbalanced within and across participants. We randomly selected the order of the blocks for the first participant in the first session and reversed the order for the last session. The order of the blocks for the next participant was the reverse of the previous participant. This was repeated for the following pairs of participants. The order of trials within each block was randomly selected under the constraint that each condition was presented once before any condition was repeated.

Thresholds were estimated by the same method as in Experiment 1 with the following exceptions; We used seven comparison stimuli. The two naïve participants had a step size of 3% diameter difference, and three of the expert participants redid three or four pair-wise comparisons with a step size of 3% or 4% diameter difference.

3.1.4. Procedure

The task and the timeline of this experiment's procedure were the same as the simultaneous presentation condition of the mean size discrimination in Experiment 1 except that the distributions varied across the blocks. The five experienced participants were given feedback after each trial, whereas the two new participants were given feedback only in the practice blocks.

3.2. Results and discussion

The results of Experiment 3 are shown in Fig. 5. We first compared within- and between-distribution pairs. The thresholds for mean discriminations within the same distributions were around 8%, which is similar to the threshold for the simultaneous condition in Experiment 1. The thresholds for mean discriminations across different distributions were around 10%. The difference was small but significant ($F_{(1,4)} = 61.464$, $p < 0.01$).

An ANOVA on the ten pairs tested showed a significant overall effect of distribution type ($F_{(9,36)} = 10.729$, $p < 0.01$). According to a Bonferroni post hoc analysis, there were no significant differences between judgments on any pairs drawn from within the same distributions, or between judgments on any pairs drawn from two different distributions, with one exception: pairs from two homogeneous distributions gave significantly lower thresholds than pairs drawn from two normal distributions. The homogeneous pairs gave the lowest threshold, which differed significantly from all the judgments between two different distributions. The judgment on a two-peaks and a homogeneous pair gave the highest threshold, which differed significantly from all judgments on pairs from the same distributions.

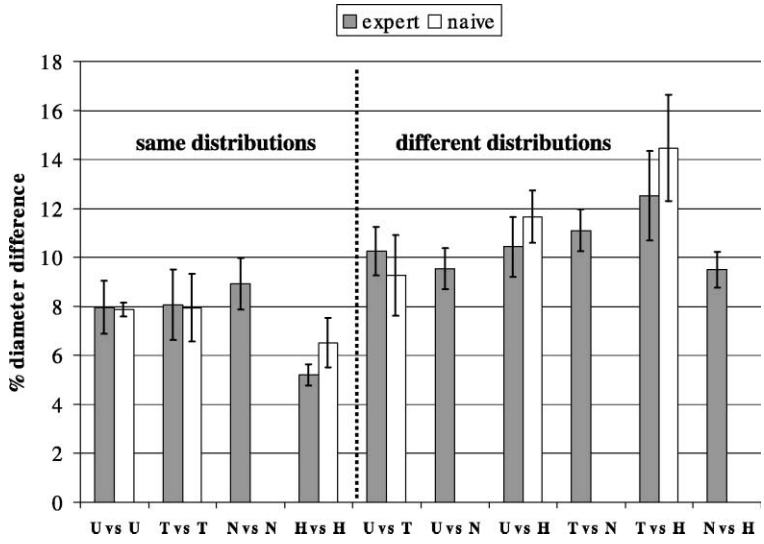

Fig. 5. The results of Experiment 3. U stands for the uniform distribution, T stands for the two-peaks distribution, N stands for the normal distribution, and H stands for the homogeneous distribution.

The fact that thresholds for discriminating the mean size between different distributions were only 2% higher than thresholds for discriminating displays from the same distribution is a critical observation for the claim that participants were indeed averaging sizes. In most cases, when the distributions are different, participants are forced to compare the means rather than any individual items. The result confirms that at least in these conditions the displays are being statistically analyzed and compared. The highest threshold involved a comparison across the two-peaks and the homogeneous displays. These are the two that differ most in appearance, with no shared sizes at all and maximally different variances. Again the fact that thresholds were only marginally higher here, at least for the experienced participants, confirms that participants are able to respond to the mean of two sizes almost as accurately as to a single size.

Thresholds for the naïve participants did not differ significantly from those of the experienced participants. The naïve participants did not get feedback during the experimental blocks, whereas the experienced participants did. These results imply that people can accurately average sizes without any period of extensive learning.

4. General discussion

The first two experiments measured thresholds for discriminating the mean sizes of two displays, comparing simultaneous with successive presentations and heterogeneous with either homogeneous multi-item displays or single item displays, which did not require any averaging process. The results were surprising. The mean judgments with heterogeneous displays were either as accurate, or close to as accurate, as the single item judgments. There was little effect on mean judgments of either the delay with successive rather than simultaneous presentation (over a range of 0–2 s) or exposure duration (over a range of 50–1000 ms). The thresholds did rise significantly with delay, but only to 10%, and with decreased presentation time but only to 8%. The increase in thresholds was if anything smaller than those for the single items. Judgments of the mean size of heterogeneous displays seem to be made both efficiently and in parallel.

Although thresholds were similar across all conditions, there were some differences that reached significance. They can be summarized as follows: first, in both of the more difficult conditions, those with brief exposures and those with long delays, the homogeneous displays gave better performance than either the heterogeneous or the single item displays. Thus the redundant presentation of multiple identical circles appears to help participants when the conditions impose extra demands either on processing speed or on memory. Secondly, the single item displays improved more than the heterogeneous displays as the exposure duration increased and as the delay was reduced or eliminated. There may be internal noise in the averaging process that sets a ceiling on the improvement that is possible with heterogeneous displays.

Thresholds in the present experiment increased only by 2% for the mean judgments as the exposure duration decreased by a factor of forty (from 2 s to 50 ms). Even allowing for some use of iconic memory, it is unlikely that any serial process of adding each size and dividing by the number of circles could be implemented. Performance was as good at 50 ms for the mean judgments as for the single circles. This highly accurate performance with such a brief exposure is consistent with the hypothesis of a separate parallel mechanism operating on sets of items to extract their mean size, and perhaps other statistical measures such as their range or variance. It may also represent statistical measures on other dimensions besides size, such as orientation, speed and direction of motion, color and other properties.

The results of Experiment 3 support our belief that the participants really were averaging sizes when they made mean size judgments. Tests involving different distributions can rule out strategies bypassing the averaging process. For example, comparisons of homogeneous displays to displays with two-peaks cannot depend on matching individual circles, since no identical stimuli are present across the pairs of displays. Yet most between-distribution thresholds were within 1% of 2% of the corresponding within-distribution thresholds and the largest difference was only 4%.

The idea that the visual system generates statistical measures of the features present in a scene was proposed in a different context by Treisman and Gormican (1988) who linked it to parallel processing in feature search tasks. Studies of visual attention (e.g. Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989), have shown a limited mental capacity for search tasks involving anything more complex than separate, highly discriminable features. This makes it important for the visual system to reduce the processing load by summarizing redundant information. Treisman (1991) and Treisman and Gormican (1988) suggested that preattentive processes pool feature information within each of a set of coarsely coded feature maps, giving an average measure of the degree to which each of these feature values is present in the display. Popout performance depends on global attention to the display as a whole. A unique target is detected if it generates activity in a set of detectors that are not also activated by the distractors. Search asymmetries arise when a single feature in which target and distractors differ is present in one of the two and absent or reduced in the other. For example, lines may be represented by their orientation and their degree of curvature. A curved line has some curvature, whereas a

straight line has none. The presence of activity in curvature detectors allows a curved line to pop out among straight ones, but not the reverse. When the target has no unique feature and activates the same detectors as the distractors, but to a lesser degree, an attention window of adjustable size is narrowed sufficiently to isolate pooled samples whose averaged signal differs detectably when the target is present in the sample and when it is not. Thus feature activity is averaged within the window of attention, allowing comparisons of feature activity within the attended area with that present in the rest of the display. The less discriminable the target is from the distractors, the more narrowly focused the attentional scan must be. Combined with the idea of coarse coding of features by ratios of activity in a few populations of detectors, this idea proved useful in explaining search asymmetries and the continuum of increasing search slopes with decreasing feature discriminability. Consistent with these suggestions, Chong and Treisman (2000) showed that statistical processing is more readily combined with global or distributed attention than with local or focused attention. The present research confirms that averaged information does become rapidly available for sets of items that are processed in parallel.

Statistical processing does not depend on conscious access to the individual items to be averaged. Crowding in the visual periphery, a form of attentional overload, can eliminate perception of particular individual items (He, Cavanagh, & Intriligator, 1996). However, Parkes et al. (2001) showed that humans could reliably estimate the average orientation even in conditions in which they were unable to report the orientation of any individual patch. Again this suggests preattentive averaging of feature information.

How might we form representations of mean values on various feature dimensions? One alternative would be to record all the individual values and average them. Parkes et al. (2001) applied an averaging model to orientation judgments. They made the additional assumption that Gaussian noise is added both in encoding the individual values and in averaging them. Their model simulated human performance quite accurately. The idea that perception of the mean depends on first registering all the individual elements is also consistent with the physiological finding that the global perception of the average direction of motion is severely impaired when cats lose a large proportion of their directionally selective neurons as a result of being reared in a restricted environment for the first 8 months of life, greatly reducing the number of directionally selective neurons (Pasternak, Albano, & Harvitt, 1990).

A simple averaging model, however, cannot fully explain our findings in mean size judgments. It would predict the same performance across distributions as within distributions, since it uses the same averaging algorithm and adds the same early and late noise to independently encoded values. Yet our size thresholds were significantly higher when the distributions were different. Judgments of the perceptual mean may be harder to abstract across differences in the range or individual elements.

Another possible mechanism might be to take a fixed sample of individual values and to average those. However this would predict decreasing accuracy as the display size increases and any given sample becomes less representative of the whole. Yet Ariely (2001) found no effect of display size, suggesting parallel registration of the whole display.

The shape of the population response across individual neurons may offer an alternative to the averaging model. If the visual system registers the distribution across individual values, it could take the peak value after normalization as representing the mean. In the domain of motion perception, Treue, Hol, and Rauber (2000) used a related idea to predict perceptual segregation of independently moving surfaces. When the distribution is too broad to be interpreted as a single direction of motion, they suggest that the perceived directions represent the activation peaks of the smallest number of Gaussian shaped activity profiles that could be summed to produce the observed activity profile. They recorded the neural responses in macaque area MT to dot patterns sliding transparently across one another, which are normally perceived as independently moving surfaces. The stimuli contained two directions. Segregation did not depend on the presence of two most strongly activated values. Rather, the visual system seemed to use the overall shape of the population response to determine the number and directions of motion components, as if the center of each Gaussian was used to represent an underlying population perceptually. Their approach explained a number of phenomena, including susceptibility of the motion system to direction metamers, where motion patterns combining three of five directions were incorrectly perceived by subjects as comprising only two directions.

An equivalent model in the size domain could explain our finding that the accuracy of mean discrimination was slightly reduced when the distributions differed, especially when one of the two was the two-peaks distribution. In the two-peaks distribution the separation between the two circle sizes was larger than in any of the other distributions. This may have resulted in occasional representation by two inferred Gaussians, and no representation of the mean.

There are many ways in which representing the statistical properties of a display may be helpful in everyday life. First accurate representation of statistical properties can help us to distinguish different surfaces by their texture, allowing us to segregate the scene into likely objects and distinct background areas-an essential step for object identification and selective attention. Julesz

(1981) found that people could preattentively distinguish texture pairs, if they had certain visual features (textons) whose first-order statistics provided the information necessary to segregate areas and establish texture borders. Nothdurft (1990, 1997) describes the statistical requirement for texture boundaries to become salient: feature variation across the boundary must be significantly greater than feature variation within the boundaries. Texture features derived from the local statistics of an image can simulate human performance (Rubenstein & Sagi, 1990) and can be used to classify satellite images (Haralick, Shanmugam, & Dinstein, 1973).

Secondly, accurate representation of the mean may facilitate detection of an odd object in a scene. Instead of comparing all objects in a scene to each other, we can compare each object to the mean and standard deviation of the background population, allowing faster detection of any outliers.

Finally, statistical representation helps to economize on the limited capacity of the visual system. Rather than preserving all the detailed information in a scene, we can abstract the statistical properties and then at retrieval fill them in using the stored statistics. Given the complexity of a typical visual scene and our limited capacity for perceiving and storing the details, we have little alternative to using summary representations.

Appendix A. Experiment to assess the perceived size of the mean of two circles or lines

A.1. Method

The stimuli were presented on the screen of a Samsung SyncMaster 955DF 19 in. Monitor, driven by a Macintosh G4, which also performed all timing functions and controlled the course of the experiment. Participants (13 Princeton undergraduates) viewed the screen with both eyes and were seated approximately 66 cm from the screen. Each display contained two circles or two lines to be averaged and one circle or line to be adjusted to match the perceived mean of the other two. The adjustable circle or line was presented in the center of the lower visual field. The other circles were presented in the center of the left and right upper visual field. The range of sizes was from 5.05° to 14.44° (diameters for the circle and lengths for the line). In each trial all of the circles and lines either remained same or were scaled by multiplying the sizes by 1.3. The same factor scaled all circles and lines in one trial. The luminance of the stimuli was 49.93 cd/m^2 and the luminance of the black background was 0.006 cd/m ^{2}.

There were three independent variables in the experiment, all of which were varied within participants. One was the type of test (either perception or memory) which was varied between blocks. The other two, which were varied within blocks, were the stimulus type (either circle or line), and the initial size of the adjustable stimulus (requiring either ascending or descending size adjustments). The initial size was randomly selected over a range of 3.60°–5.01° in ascending trials and 15.89°–14.48° in descending trials. Each block started with two practice trials, followed by 48 trials (2 stimulus types × 2 initial sizes of the adjustable stimuli × 2 multiplicative factors × 6 repetitions). The order of blocks was counterbalanced across participants. The order of trials within each block was randomly selected under the constraint that each condition was presented once before any condition was repeated.

In the perception block, two stimuli and an adjustable stimulus were presented until participants completed their adjustments. Participants were asked to set the adjustable circle to match the estimated mean size of the two circles. They could decrease the size of the adjustable circle by 0.49°, whenever they pressed '1'. They could increase the size of the adjustable circle by the same amount, whenever they pressed '2'. When they finished their adjustments, they could move on to the next trial by pressing '9'. In the memory block, the procedure was the same as in the perception block except that the two stimuli disappeared after 1 s. The adjustable circle was present from the beginning in the memory block.

A.2. Results and discussion

The results are shown in Fig. 6. The mean size estimates did not differ significantly in the perception and memory conditions ($F_{(1,84)} = 2.996$, $p = 0.09$), but the variance of the size estimates was significantly larger in the memory condition than in the perception condition ($F_{(1,84)} = 15.192$, $p < 0.01$), suggesting some decrease in accuracy over time. The estimated mean size was larger for the lines than for the circles ($F_{(1,84)} = 10.643$, $p < 0.01$) and the variance of the line-size estimates was also larger than that of the circle-size estimates ($F_{(1,84)} = 7.021$, $p < 0.01$). No two-way or three-way interactions were significant. Since the other main effects did not vary with the size of the set, we averaged the data of the larger set and the smaller set.

The left side of Fig. 6 shows the presented sizes and the possible mean sizes according to different calculation methods. Participants' estimates differed significantly from the geometric mean ($t_{(25)} = 16.315$, $p < 0.01$), the arithmetic mean of the diameters ($t_{(25)} = 4.762$, $p < 0.01$), and the arithmetic mean of the areas ($t_{(25)} = -5.514$, $p < 0.01$).

Teghtsoonian (1965) investigated judgments of size using the method of magnitude-estimation. She found that the judged size of a circle was related to its area by a power function with an exponent of 0.76. In order to see whether this formula would also predict our data on

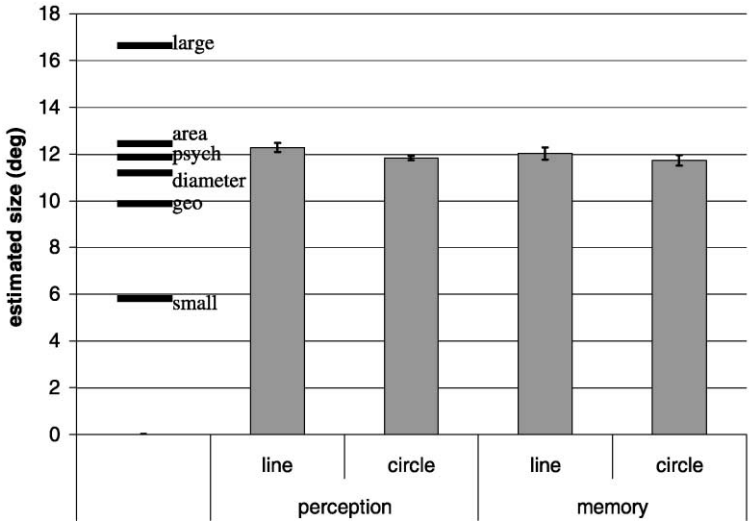

Fig. 6. The judged mean sizes. The bars on the left indicate the presented sizes (small and large), the geometric mean (geo), the arithmetic mean of the diameters (diameter), the mean of the areas on a power function with exponent of 0.76 (psych), and the arithmetic mean of the areas (area).

perceptual mean sizes, we converted the areas of the two presented sizes in our experiment using a power function with an exponent of 0.76, calculated the mean size of the two predicted sizes, and then converted the mean back into a physical size. This predicted perceptual mean size was a good approximation of the participants' estimates and did not differ from them statistically ($t_{(25)} = -0.871$, $p = 0.39$) either for the perception ($t_{(12)} = -0.472$, $p = 0.65$) or for the memory conditions ($t_{(12)} = -0.728$, $p = 0.48$). Note that the power function with the exponent of 0.76 predicts a mean that lies between the means of the areas and the means of diameters. One possible explanation of our results is that participants divided their estimates between matching the mean area and matching the mean diameter length. The values are probably too close for our data to distinguish whether the participants could be divided into two groups, one matching each of those criteria. The same kind of compromise also had determined the size judgments made by Teghtsoonian's observers. She instructed one group of participants specifically to judge size on the basis of area and found an exponent of 1.03. When they were given no particular instructions, the exponent dropped to 0.76, consistent with a mixture of judgments based on area and judgments based on diameter.

In the case of the lines, our participants' estimates showed a similar bias, giving an estimate of mean length that was significantly larger than the arithmetic mean ($t_{(25)} = 5.817$, $p < 0.01$). Our results differ from those of Teghtsoonian (1965), whose participants gave judged sizes related to length by a power function with an exponent of 0.98, which was not significantly different from 1. In Teghtsoonian's experiments lines and circles were blocked, whereas they were intermixed in our experiment. The estimates of the circle sizes in our mixed blocks may have influenced estimates of line length.

References

Ariely, D. (2001). Seeing sets: representation by statistical properties. *Psychological Science, 12*, 157–162.

Ariely, D., & Burbeck, C. A. (1995). Statistical encoding of multiple stimuli: a theory of distributed representation. *Investigative Ophthalmology and Visual Science, 36*(Suppl.), 8472 (Abstract).

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Brenner, N., Bialek, W., & de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron, 26*, 695–702.

Chong, S. C., & Treisman, A. (2000). *Effects of divided attention on the representation of a visual scene*. Paper was presented at '00 OPAM, New Orleans.

Dakin, S. C. (1997). The detection of structure in glass patterns: psychophysics and computational models. *Vision Research, 37*, 2227–2246.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research, 37*, 3181–3192.

Davidson, M. L., Fox, M. J., & Dick, A. O. (1973). Effect of eye movements on backward masking and perceived location. *Perception & Psychophysics, 14*, 110–116.

De Bruyn, B., & Orban, G. A. (1988). Human velocity and direction discrimination measured with random dot patterns. *Vision Research, 28*, 1323–1335.

Finney, D. J. (1971). *Probit analysis*. Cambridge, UK: Cambridge University Press.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics, 3*, 610–621.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383*, 334–337.

Heeley, D. W., & Buchanan-Smith, H. M. (1990). Recognition of stimulus orientation. *Vision Research, 30*, 1429–1437.

Hochberg, J. E. (1978). *Perception*. Englewood Cliffs, NJ: Prentice Hall.

Hock, H. S., & Schmelzkopf, K. F. (1980). The abstraction of schematic representations from photographs of real-world scenes. *Memory & Cognition, 8*(6), 543–554.

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology, 23*, 420–456.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290*, 91–97.

Müller, J. R., Metha, A. B., Krauskopf, J., & Lennie, P. (1999). Rapid adaptation to visual cortex to the structure of images. *Science, 285*, 1405–1408.

Nothdurft, H. C. (1990). Texton segregation by associated differences in global and local luminance distribution. *Proceedings of the Royal Society of London, B 239* (pp. 295–320).

Nothdurft, H. C. (1997). Different approaches to the encoding of visual segmentation. In L. Harris & M. Jenkins (Eds.), *Computational and psychophysical mechanisms of visual segmentation* (pp. 20–43). New York: Cambridge University Press.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*, 739–744.

Pasternak, T., Albano, J. E., & Harvitt, D. M. (1990). The role of directionally selective neurons in the perception of global motion. *The Journal of Neuroscience, 10*, 3079–3086.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science, 8*(5), 368–373.

Rubenstein, B. S., & Sagi, D. (1990). Spatial variability as a limiting factor in texture discrimination tasks: implications for performance asymmetries. *Journal of Optical Society of America A, 7*, 1632–1643.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., & Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature, 386*, 69–73.

Snowden, R. J., & Braddick, O. J. (1991). The temporal integration and resolution of velocity signals. *Vision Research, 31*(5), 907–914.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153–181.

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78*, 392–402.

Treisman, A. (1991). Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 652–676.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 16*, 97–136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review, 95*, 15–48.

Treue, S., Hol, K., & Rauber, H.-J. (2000). Seeing multiple directions of motion—physiology and psychophysics. *Nature Neuroscience, 3*, 270–276.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research, 32*, 931–941.

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: the integration of direction information. *Vision Research, 29*, 47–59.

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Research, 24*, 55–62.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model of visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 419–433.